

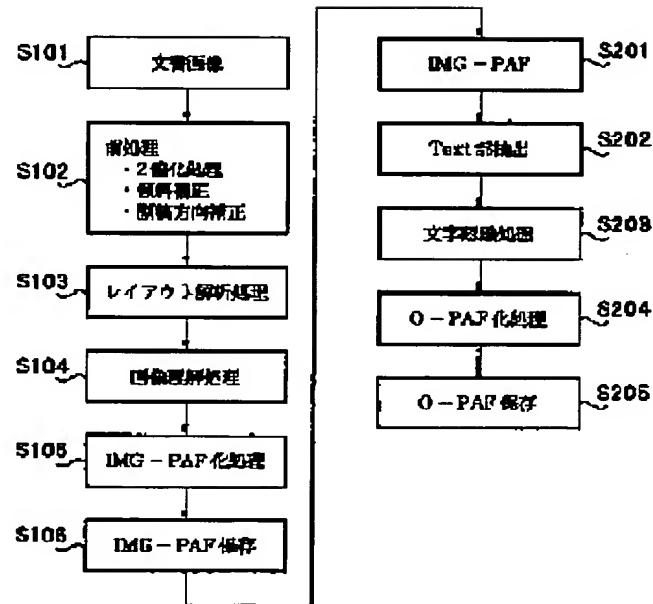
INFORMATION PROCESSOR AND METHOD THEREFOR

Patent number: JP2001076095
Publication date: 2001-03-23
Inventor: TAKAOKA MAKOTO
Applicant: CANON KK
Classification:
 - international: G06K9/20; H04N1/40
 - european:
Application number: JP19990251932 19990906
Priority number(s): JP19990251932 19990906

[Report a data error here](#)

Abstract of JP2001076095

PROBLEM TO BE SOLVED: To prepare electronic document data which suitably saves and delivers a document image and are easily electronically documented. **SOLUTION:** The layout of an inputted document image is analyzed (S103) to recognize a texture area and a picture area. The overlap area where the recognized text area and picture area overlap with each other is detected (step S104), and the storage style of the document image is switched, according to whether the overlap area is detected, thereby storing document image data (S105 and S106).



Data supplied from the **esp@cenet** database - Worldwide

特開2001-76095

(P2001-76095A)

(43)公開日 平成13年3月23日(2001.3.23)

(51)Int.C1.⁷ 識別記号
 G 06 K 9/20 3 4 0
 H 04 N 1/40
 1/41

F I
 G 06 K 9/20 3 4 0 L 5B029
 H 04 N 1/41 Z 5C077
 1/40 F 5C078

審査請求 未請求 請求項の数 13 O L

(全 10 頁)

(21)出願番号 特願平11-251932

(22)出願日 平成11年9月6日(1999.9.6)

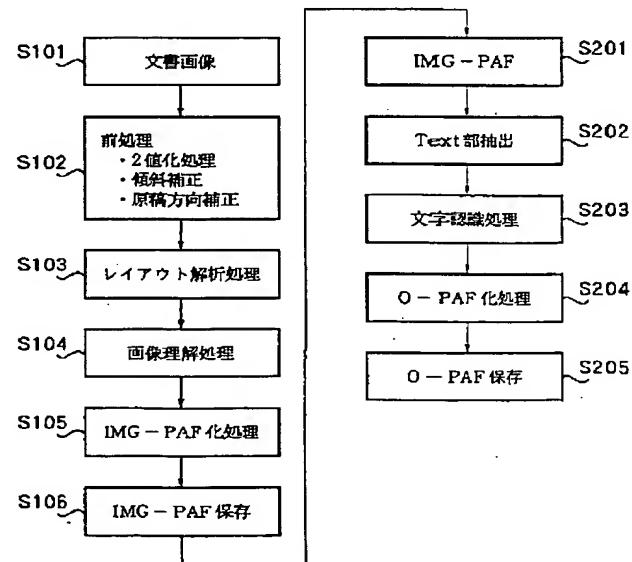
(71)出願人 000001007
キヤノン株式会社
東京都大田区下丸子3丁目30番2号(72)発明者 高岡 真琴
東京都大田区下丸子3丁目30番2号 キヤノン株式会社内(74)代理人 100076428
弁理士 大塚 康徳 (外2名)
F ターム(参考) 5B029 BB02 CC29 DD10 EE04
5C077 MP06 MP08 PP27 PQ08 RR02
RR21
5C078 AA04 AA09 CA03

(54)【発明の名称】情報処理装置及びその方法

(57)【要約】

【課題】文書画像の保存や配信に適し、しかも電子文書化することが容易な電子文書データを作成することを可能とする。

【解決手段】入力された文書画像についてレイアウト解析を行い(S103)、テキスト領域とピクチャ領域とを認識する。そして、認識されたテキスト領域とピクチャ領域とが重なる重複領域を検出し(ステップS104)、重複領域が検出されたか否かに基づいて当該文書画像の格納形態を切り替えて該文書画像データを格納する(S105、S106)。



【特許請求の範囲】

【請求項1】 入力された文書画像についてレイアウト解析を行い、少なくともテキスト領域とピクチャ領域とを認識する解析手段と、前記解析手段で認識されたテキスト領域とピクチャ領域とが重なる重複領域を検出する検出手段と、前記検出手段で重複領域が検出されたか否かに基づいて当該文書画像の格納形態を切り替えて該文書画像データを格納する格納手段とを備えることを特徴とする情報処理装置。

【請求項2】 前記格納手段は、前記検出手段で重複領域が検出されなかった場合は、前記検出手段で検出されたテキスト領域とピクチャ領域とを別個の部分画像として格納することを特徴とする請求項1に記載の情報処理装置。

【請求項3】 前記格納手段は、前記検出手段で検出された重複領域を一つのピクチャ領域として扱うことを特徴とする請求項1に記載の情報処理装置。

【請求項4】 前記格納手段は、テキスト領域とピクチャ領域の夫々に適当な圧縮処理を施して前記文書画像データを格納することを特徴とする請求項2または3に記載の情報処理装置。

【請求項5】 前記格納手段は、前記検出手段で検出された重複領域の前記文書画像に占める割合が所定閾値よりも大きい場合、当該文書画像を部分画像に分けずに格納することを特徴とする請求項1に記載の情報処理装置。

【請求項6】 前記テキスト領域として格納された領域の部分画像について文字認識処理を実行する文字認識手段を更に備えることを特徴とする請求項1に記載の情報処理装置。

【請求項7】 入力された文書画像についてレイアウト解析を行い、少なくともテキスト領域とピクチャ領域とを認識する解析工程と、前記解析工程で認識されたテキスト領域とピクチャ領域とが重なる重複領域を検出する検出工程と、前記検出工程で重複領域が検出されたか否かに基づいて当該文書画像の格納形態を切り替えて該文書画像データを格納する格納工程とを備えることを特徴とする情報処理方法。

【請求項8】 前記格納工程は、前記検出工程で重複領域が検出されなかった場合は、前記検出工程で検出されたテキスト領域とピクチャ領域とを別個の部分画像として格納することを特徴とする請求項7に記載の情報処理方法。

【請求項9】 前記格納工程は、前記検出工程で検出された重複領域を一つのピクチャ領域として扱うことを特徴とする請求項7に記載の情報処理方法。

【請求項10】 前記格納工程は、テキスト領域とピクチャ領域の夫々に適当な圧縮処理を施して前記文書画像

データを格納することを特徴とする請求項8または9に記載の情報処理方法。

【請求項11】 前記格納工程は、前記検出工程で検出された重複領域の前記文書画像に占める割合が所定閾値よりも大きい場合、当該文書画像を部分画像に分けずに格納することを特徴とする請求項7に記載の情報処理方法。

【請求項12】 前記テキスト領域として格納された領域の部分画像について文字認識処理を実行する文字認識工程を更に備えることを特徴とする請求項7に記載の情報処理方法。

【請求項13】 文書画像データの格納処理をコンピュータに実現させるための制御プログラムを格納する記憶媒体であって、該制御プログラムが、

入力された文書画像についてレイアウト解析を行い、少なくともテキスト領域とピクチャ領域とを認識する解析工程のコードと、

前記解析工程で認識されたテキスト領域とピクチャ領域とが重なる重複領域を検出する検出工程のコードと、前記検出工程で重複領域が検出されたか否かに基づいて当該文書画像の格納形態を切り替えて該文書画像データを格納する格納工程のコードとを備えることを特徴とする記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、光学的読み取り手段によって読み取った文書原稿を扱い、そのデータを再利用する情報処理装置及び方法に関する。

【0002】

【従来の技術】 従来、文書画像を保存、配信する場合、白黒2値画像ならばMMR圧縮を用い、カラー画像ならばJPEG圧縮を用いて、画像を保存したり配信したりしてきた。さらに近年では、画像認識技術が向上し、単に画像を保存、配信するだけでなく、電子文書として再利用する要望も高まり、実現されつつある。

【0003】 通常、文書画像を再利用する場合、認識処理に適した画像形式に変換し、レイアウト解析処理を行い、文字部については文字認識処理を行い、電子文書化する。しかしながら、レイアウト解析処理や文字認識処理の両方ともに判断ミスを犯すことが間々ある。特に文字認識処理は、99%の精度であっても残りの1%のミスはどうしても起きてしまうことになる。これは、単に文字認識処理が原因ではなく、前処理である画像2値化処理やレイアウト解析処理が原因であったりする場合もある。いずれにせよ完全に再現することは不可能に近い。そのため、従来から行われている画像圧縮を用いて、画像情報として保存や配信をすることが一般的に行われている。

【0004】

【発明が解決しようとする課題】 しかしながら、従来の

手法において、例えばカラー文書画像を保存や配付する場合には、原画像を比較的忠実に圧縮するJPEG圧縮がよく用いられる。JPEG圧縮画像であると、例えば受信した側の装置が内容を確認するために、伸長処理をすることになるが、その際、原画像が高解像度であったりサイズが大きかったりすると確認するのに多くのメモリが必要となり、また、伸長処理に費やされる時間も多くなってしまう。

【0005】そのため、複数ページにわたる文書画像の場合、受信した側は、内容を確認することなく、データを廃棄してしまうこともあった。

【0006】そこで、カラー文書画像を、ファイルサイズを小さくし、伸長処理の際に必要なメモリ容量を低減するため、白黒2値化して、それを配信することが行われている。しかしながら、このように処理すると、カラー情報の欠落や2値化時に写真部が黒くつぶれたり、あるいはかすれたりし、画質が著しく劣化することになる。更に下地付きの文書画像の場合、ノイズっぽい画像になってしまることがある。白黒2値画像を用いるのは、結局文字部が文書の中で重要なため、文字部がはっきり読めればよいという妥協の上で用いられる。よい例がFAXで送る例である。FAXは、最終的には、白黒文書でしか扱わないものであり、多くの場合はそれで事足りている。

【0007】白黒2値画像の場合、代表的なものとして、単純2値化された画像と誤差拡散によって2値化された画像を扱う場合が多い。前者だと文字部はきれいだが、写真部は、黒くつぶれてしまう。一方、後者だと、MMR圧縮ではファイルサイズが大きくなり、文字部には、バックノイズが出現してしまう。殊に、後者の場合、上述のような再利用を目的として電子文書化するのはかなり難しい。

【0008】いずれにせよ、画像を扱う場合には、その文書画像をコードとして扱える電子文書化するにはかなり障壁が高かった。ましてや、再利用することはあまり行われないのが実状であった。

【0009】そこで、以下のような要件を満たす電子文書形式が望まれていた。・カラー画像はカラーのまま、小さいサイズとなり、しかも再生画像が、きれいであること。・下地のある白黒画像であっても、バックノイズが出現して、醜くないように、必要なところははっきりとさせ、不必要なところは、削除されたきれいな再生画像であること。・そして、いずれもコンピュータ上のアプリソフトに取り込める再利用可能な電子文書であること。

【0010】従って、本発明の目的は、上述の如き要件を満たす電子文書形式を生成することが可能な情報処理装置及び方法を提供することにある。

【0011】すなわち、本発明は、文書画像の保存や配信に適して、しかも電子文書化することが容易である文

書画像データを作成可能な情報処理方法及び装置を提供することを目的とする。

【0012】

【課題を解決するための手段】上記の目的を達成するための本発明による情報処理装置は例えば以下の構成を備える。すなわち、入力された文書画像についてレイアウト解析を行い、少なくともテキスト領域とピクチャ領域とを認識する解析手段と、前記解析手段で認識されたテキスト領域とピクチャ領域とが重なる重複領域を検出する検出手段と、前記検出手段で重複領域が検出されたか否かに基づいて当該文書画像の格納形態を切り替えて該文書画像データを格納する格納手段とを備える。

【0013】

【発明の実施の形態】以下、添付の図面を参照して本発明の好適な実施形態を説明する。

【0014】本実施形態では、PAF (Page Analysis Format) と呼ぶ電子文書フォーマットを仮に設定する。このPAFを仲介として、前記課題を解決する。

【0015】<PAF>PAF (パフ) では、文書画像を画像解析し、文書画像中のコード化が可能な領域は、できる限りコード化させる。コード化が難しい領域は、部分画像で保存する。図1は本実施形態による電子文書フォーマット (PAF) を説明する図である。11はPAF全体構成を示し、図示を省略したが、文書全体のプロパティ情報 (例えば、何ページの文書PAFか等を管理する項目) や、管理情報 (例えば、何時にどのようなスキーナで入力したものか) 等の情報を格納する領域を有する。

【0016】Text Object部12は、文書画像中のText領域と判断した領域を保持する。Text領域は、最も圧縮効率が高い領域である。このText領域には

30 2つの保存方法でデータが保持される。1つはText領域を2値画像で保持する方法である。これは、白黒2値画像は当然のことながら、RGBフルカラー画像もText領域に限っては、2値画像で保持する。一方、もう一つの保存方法は、Text領域を文字認識処理を実行して文字コード化して保持する方法である。文字認識コード化されたText領域は再利用可能な文書画像となる。

40 【0017】Picture Object部13は、写真などの自然画や解析ではまだベクトルコード化ができない画像領域を部分画像で保存する。ここでは、原画像が2値画像の場合は、部分2値画像で保持し、多値画像の場合は、部分多値画像で保持する。

【0018】Drawing Object部14は、線や線画、囲い枠などの線ベクトル化が可能な領域を保持する部分である。例えば、横線の場合、その開始位置、終了位置、太さ、線種などを保持する。

50 【0019】Table Object部15は、文書画像中の表領域と判断された領域を保持する部分である。例えば、表

の縦横のセル数や、そのセル構成情報、表枠情報の他、セル内部に書かれた Text も保持する。

【0020】文書画像をレイアウト解析し、その解析結果に基づき、以上のような形式で保存した電子文書フォーマットを P A F と呼ぶ。

【0021】図2は、各オブジェクトごとに適切な圧縮を行い保存する適応的圧縮を示す図である。上述したように、画像として保持する場合、その領域に適した圧縮方法を選択して圧縮、格納する。これによりファイルサイズが小さくなる。上記のText Object領域12とTable Object領域15は、2値画像として保持する場合がある。この場合、線を示すDrawing Object部14を除きすべて部分画像の集まりとなる。後述の説明を分かり易くするため、以上のような形式の保存を I M G - P A F (イメージパフ) と呼ぶ。また、レイアウト解析処理結果より、Text を含む画像領域と判断された領域に対して、文字認識処理を行った結果を保存した形式を O - P A F と呼ぶ。

【0022】以上説明したような電子文書 P A F を文書画像の代わりに保存することにより、ファイルサイズを縮小したり、再利用を可能としたりすることができる。また、その他、全部文検索を可能とするなどの利点も生まれる。

【0023】しかしながら、このような P A F 化をするにあたり、どのような文書画像も適切に部分オブジェクト化できるわけではない。図2に示した文書画像は比較的整然とレイアウトされた文書であるため、P A F 化しても効率よく縮小ファイル化できる。また再現文書画像も忠実に行なうことが可能である。ところが、もっと複雑に入り組んだ文書画像の場合、そのままレイアウト解析した結果の I M G - P A F を保存していても重複領域の処理が複雑化して、かえって使い物にならない再生電子文書となってしまうことがある。この点に関して、図4～図6を用いて説明する。

【0024】図4は、レイアウト解析の結果、各領域が整然と配置された文書画像を示す図である。図5は、レイアウトが少し入り組んだ文書画像を示し、図6は、P I C T 部(イメージ領域)に T E X T 部(テキスト領域)が重なったり、中に入ったりする文書画像レイアウト解析結果を示す。図6のような複雑な解析結果をそのまま部分領域画像で保存した場合、重複する部分が多く、返ってファイルサイズが大きくなってしまったり、また再生する際も、画像中の T e x t 部が変に浮き出てしまうといったような再現画像となってしまう。すなわち、どの文書画像は部分保存方法でよく、どの文書画像では返って不効率かつ見苦しい画像になってしまうかをあらかじめ判断することが必要である。

【0025】本実施形態では、図4のような文書画像の場合は極力細分化した部分画像として扱い、図6のような複雑な文書画像の場合は、再生を考慮して、ある程度

大枠の部分画像として扱うようにする分類技術を加えることにより、どのような文書画像も P A F 化して、その利点を生み出すようにする。さらに、各部分画像毎の保存を行う部分切り出し画像保存の処理では、うまく再生できないような文書画像は、原画像保存のままでよいと判断する。これにより、レイアウト解析や文字認識が悪い結果を出しそうな場合、できるだけ原画像の情報を保存するようにでき、そうでない良好な文書画像は、できるだけ圧縮効率を上げ、再利用可能に文書を電子化することができる。

【0026】以下、上記 P A F 化までの本実施形態の処理について説明する。図12は本実施形態による文書画像処理を実行するシステムの構成を示すブロック図である。図12において、101はCPUであり、R O M 1 0 2 や R A M 1 0 3 に格納された制御プログラムを実行して、各種処理を実現する。104は外部記憶装置であり、C P U 1 0 1 による実行のために適宜 R A M 1 0 3 ロードされる各種アプリケーションプログラムや、文書画像データを格納する。105はディスプレイであり、C P U 1 0 1 の制御の下で各種の表示を行う。106は操作部であり、キーボードやポインティングデバイス等を具備する。107はバスであり、上述の各構成を接続する。なお、図示のフローチャートを参照して後述される処理を実現するための制御プログラムは、外部記憶装置104に格納され、必要に応じて R A M 1 0 3 にロードされて C P U 1 0 1 により実行されるものとする。もちろん、これら制御プログラムは、R O M 1 0 3 に格納されていてもよい。

【0027】図3は、本実施形態による P A F 化処理の手順を説明するフローチャートである。まず、ステップ S 1 0 1 で文書画像を入力する。ここで入力される文書画像とは、スキャナ画像もしくは、画像ファイルである。ステップ S 1 0 2 では前処理を行う。前処理においては、もし当該文書画像が多値画像であればこれに対して2値化処理を行う。例えば、R G B カラー画像の場合は、まず最初に白黒グレースケール変換してから2値化処理を行う。この2値化処理は、プリントのための2値化ではなく、後のレイアウト解析処理のための2値化処理である。例えば128をスライスレベルとした単純2値化処理方式や、文書画像中の部分領域でヒストグラム解析した結果で適切な閾値を求め、それをスライスレベルとして2値化処理を行う方式等で2値画像を作成する。この2値化処理を用いる理由は、後のレイアウト解析処理や文字認識処理では2値画像を用いるのが一般的だからである。さらに付加的な理由として、下地の色を確実に取り除くためである。下地の中の文字は、領域判別、文字きり出しが判断しにくいからである。一方 P I C T 部は一様に黒く塗りつぶされた領域としていた方がよい。黒が飛んでいたりすると、P I C T 部の中に T E X T 部があるような誤認識しやすいからである。

【0028】さらに、文書画像が傾いていると認識ミスを犯しやすい。そこで、文書画像の傾斜補正を行う。また文書画像が左を向いていたり、さかさまを向いていることがないように、原稿方向補正処理も行う。

【0029】ステップS103へ進み、レイアウト解析処理を行う。レイアウト処理の処理方法は、ここでは、公知技術である輪郭線追跡によるレイアウト解析等の手法を用いることができ、画像中も黒画素の塊を検出してその輪郭をたどる輪郭線追跡方法や、黒画素領域を検出したら順番に番号を付加して行くラベリング方式がある。そして、検出した黒画素の塊の大きさ、位置等からその領域の属性を判断する。

【0030】その結果、レイアウト解析処理を行うと、TEXT（文字）、TITLE（タイトル）、CAPTION（キャプション）、LINEART（線画）、PICTURE（自然画）、FRAME（枠）、LINE（線）、TABLE（表）などの属性毎に認識された各ブロックの属性情報とその矩形アドレス情報が、レイアウト解析結果として出力される。上記図1のText Object部12に属する領域は、TEXT（文字）、TITLE（タイトル）、CAPTION（キャプション）である。同様にPicture Object部13は、PICTURE（自然画）、LINEART（線画）、Drawing Object部14は、FRAME（枠）、LINE（線）、そしてTable Object部15はTABLE（表）が当てはまる。

【0031】次に、ステップS104において、画像理解処理を行う。この処理は本実施形態における重要な処理であり、その詳細は後述する。この画像理解処理の結果、文書画像は適切にPAF化を行うために分類される。

【0032】ステップS105では、IMG-PAF化処理を行う。このIMG-PAF化処理では、ステップS104における画像理解処理のルールに従って、部分画像の切り出しを行う。PICT部は、原画像が、多値画像の場合は、多値画像のまま切り出す。

【0033】次に、ステップS106において、IMG-PAF保存処理を行う。ここでは、ステップS105で切り出された画像についてそれぞれ適応した圧縮を行う。そして、レイアウト情報とともに圧縮画像も保存してIMG-PAF化する。

【0034】続いて、ステップS201において、IMG-PAF化保存されたデータを再び読み込む。ステップS202で、レイアウト情報より、Text領域を抽出する。Text領域には、2値の切り出し画像が保管されている。この画像が圧縮されている場合は伸長する。ステップS203では、Text領域で保存していた画像に対して文字認識処理を実行する。この文字認識処理では、前処理として、言語判別を行う。一般的には日英判別を行う。次に組方向判別を行い、縦書き・横書きを判別する。そして、例えば日本語ならば日本語文字

認識エンジンを用いて、処理を実行する。

【0035】ステップS204におけるO-PAF化処理では、ステップS203にて得られた文字コード及び文字認識情報を再びPAFの中に組み入れる。この場合、Text領域の画像は必要に応じて消去してもよい。そして、ステップS205におけるO-PAF保存処理では、ステップS204で得られたO-PAFを保存する。

【0036】以上のようなフローを実現するソフトウェアを、以下、PAF Captureと呼ぶ。さて、本実施形態のポイントは、ステップS103によるレイアウト解析結果を用いて、文書画像がどのような分類に属し、どのような画像保存を行ったら良いか判断する。この処理を画像理解処理と呼ぶ。

【0037】<文書画像の分類について>文書画像は、そのレイアウトに注目して分類することができる。図7は本実施形態による文書画像の分類体系を示す図である。文書画像51は、まず、整列レイアウト画像52、非整列レイアウト画像53の2種類に大きく分類される。整列レイアウト画像52とは、図4に示したような各属性の矩形が、整然とレイアウトされた文書である。例えば特許公報の文書はこの分類に属する。

【0038】非整列レイアウト文書53とは、図6に示したようなPICT領域とText領域が重なりあったり、中に含まれたりするような文書画像を指す。このように、非整列レイアウト画像53は、完全には領域分離が難しい文書画像である。

【0039】さて、非整列レイアウト画像53は、さらに、重複矩形レイアウト54とその他レイアウト55とに分離できる。図6のレイアウトは、この重複矩形レイアウト54に含まれる。図6のようなレイアウトの場合、画像切り出しルールさえ確立すれば、かなり良い再現結果を得ることが可能である。その他のレイアウト55に含まれる文書画像は、例えば、文字の無い画像や、画像が傾いていたために領域判別結果が極端に歪んだ結果を出力した場合に、ここに分類される。このその他のレイアウト55に分類された文書画像は、無理に部分領域画像にしなくて、原画像のまま残しておく方が返ってよい。

【0040】本実施形態では、整列レイアウト画像52、重複矩形レイアウト54の分類に当てはまる文書画像を抽出し、それ以外は、切り出し画像保存の処理系に提供せずに、無理な画像の切り出しを防止する。

【0041】図8はPICT部とTEXT部の重なりのパターンを示す図である。図8において、(a)はPICT部の内部にそっくりTEXT部が含まれてしまう例である。図9の(a)には、図8(a)のような矩形が得られる画像例を示した。また、図8の(b)はPICT部でTEXT部の一部が重なっている場合であり、図9の(b)はその画像例である。この2種類のパターン

は、レイアウト解析時に、P I C T部を親として、T E X T部のオーバーラップフラグを立てておく。ただし、

(a) の場合と (b) の場合とではそれぞれ異なるフラグを立てておく。図8の (c) はP I C T部の中にP I C T部があると解析した例であり、図9の (b) はその画像例である。同様に図8の (d) はP I C T部ともう一つP I C T部が一部重なっている場合を示すものであり、図9の (d) はその画像例である。

【0042】以上の様な画像例を用いて、ステップS 1 0 4の画像理解処理を説明する。図10は、本実施形態による画像理解処理の手順を示すフローチャートである。まず、ステップS 3 0 1においてレイアウト解析を行う。ステップS 3 0 2では、レイアウト解析結果にオーバーラップフラグを付加する。このオーバーラップフラグは、上記図8及び図9の (a) 及び (b) に示したように、P I C T部とT E X T部とがオーバーラップする場合に付加される。

【0043】ステップS 3 0 3では、レイアウト解析結果に対してオーバーラップフラグを検査する。P I C T部とT E X T部とでオーバーラップするところが無いならば、ステップS 3 0 4へ進み、当該画像にT E X T部が存在するかどうかを判定する。T E X T部が存在するならば、整列レイアウト画像と判断してステップS 3 0 6へ進み、ステップS 1 0 5のI M G - P A F化を実行する部分切り出し画像処理系へ送る。一方、ステップS 3 0 4において、T E X T領域が存在しなければ、他のレイアウト画像であると判定してステップS 3 1 0へ進む。

【0044】ステップS 3 0 3においてオーバーラップがあると判断された場合は、ステップS 3 0 5へ進み、P I C T部に少なくとも一部が重なるT E X T部を検出し、T E X T部のはみ出し部分をも含めたP I C T領域へと拡張する。例えば図6のP I C T部の場合には、図11のようなP I C T部に変更する。ここでは、大きな外接矩形とはせずに矩形を接続した状態にしておく。

【0045】ステップS 3 0 7では、オーバーラップした矩形以外の他の矩形が存在するか、検出する。無い場合は、ステップS 3 1 0へ進み、他のレイアウトと判断する。この文書画像の場合、P I C T部とT E X T部が複雑に入り組み、分離するより、一つの文書画像のままの方がよいと判定されるからである。

【0046】ステップS 3 0 7で、他の矩形があると判定された場合は、ステップS 3 0 8へ進み、再結合した矩形が文書画像の80%以上を占めるかどうかを判定する。この80%は、一例であり、矩形の位置、幅、高さを考慮して、判断する。ここで80%以上と判定された場合には、他のレイアウトと判断してステップS 3 1 0へ進む。この場合は、P I C T部とT E X T部が複雑に入り組んだ領域が多く存在し、分割しないほうが良いと判断されるからである。一方、80%以下であった

場合は、ステップS 3 0 9へ進み、重複矩形レイアウトと判断される。

【0047】整列レイアウトもしくは重複レイアウトと判定された文書画像は、共にステップS 1 0 5のI M G - P A F化処理における部分画像保存処理系で、部分画像保存を行う。その際、整列レイアウトの場合は問題なく、部分画像切り出しを行いそれぞれ最適な圧縮を行う。重複レイアウトの文書画像は、前記T E X TとP I C Tのオーバーラップ部分は、接続処理を行い、P I C T部とする。P I C T部とP I C T部のオーバーラップ部分は、同様に接続処理を行い、P I C T部とする。P I C T部内にT E X T部がある場合とP I C T部がある場合は、大きい方が小さい方を包含して画像保存する。

【0048】このような、画像理解ルールを用いることにより、I M G - P A F化したファイルが、不自然に部分画像とならないように、無理な場合は、元画像を自動保存させることができる。

【0049】以上の様な処理により、整列レイアウト画像の場合には、かなり小さなファイルとなる。また、この分類に属した文書画像は、さらに文字認識処理により、さらに使いやすい電子文書化が可能となる。一方、重複レイアウト画像は、できるだけ原画像部を残し、再現させる。これにより、不自然なT E X T部とP I C T部の境目をなくすことが可能となる。

【0050】【他の実施形態】次に他の実施形態について説明する。上記実施形態では、I M G - P A Fを保存する際に画像理解ルールを保存時に用いたが、文書画像自動分類技術は、画像ファーリングの自動分類方法としても実施できる。この場合、レイアウト解析した結果について画像理解処理を行い、図7に示した分類を得る。そして、その分類の情報をファイルの管理情報に記述しておく。そして、整列レイアウト画像52、重複矩形レイアウト画像54に属する文書画像は、少なくとも再利用可能なドキュメントとして分類しておく。その他レイアウト55に属する文書画像は、画像データとして扱う分類に分ける。このドキュメントは、画像として扱うこととする。

【0051】この分類分けは、ファイルを保管する領域を設定したりするのに都合がよい。また検索するのも、整列レイアウト画像に属するものから順番にアクセスする方が、ヒットする時間も早くなる。例えば、その他レイアウトの文書画像は、検索項目に入れないとすると、その分の処理が軽減されることになる。

【0052】以上説明してきたように、本実施形態の文書画像自動分類によれば、人間が、この画像は画像のままがよいとか、この画像は部分画像にして保存しておいても良いといったことを判断することを行わなくて済むことになる。また、整列レイアウト画像のように、文字認識を実行して、再びコード化して再利用しても、かなり良好な電子文書を得ることができるといったような判

断を、自動分類されたファイル類から判断することができる効果がある。

【0053】なお、本発明は、複数の機器（例えばホストコンピュータ、インタフェイス機器、リーダ、プリンタなど）から構成されるシステムに適用しても、一つの機器からなる装置（例えば、複写機、ファクシミリ装置など）に適用してもよい。

【0054】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体（または記録媒体）を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはCPUやMPU）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているオペレーティングシステム(OS)などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【0055】さらに、記憶媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張カードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張カードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行い、そ

の処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【0056】

【発明の効果】以上説明したように、本発明によれば、文書画像の保存や配信に適し、しかも電子文書化することが容易な電子文書データを作成することが可能となる。

【図面の簡単な説明】

【図1】電子文書PAFの簡単な構造を示す図である。

10 【図2】IMG-PAFの適応的圧縮を説明した図である。

【図3】本実施形態による本実施形態によるPAF化処理の手順を説明するフローチャートである。

【図4】整列レイアウト画像のレイアウト解析結果を示す図である。

【図5】入り組んだ整列レイアウト画像のレイアウト解析結果を示す図である。

【図6】重複矩形レイアウト画像のレイアウト解析結果を示す図である。

20 【図7】文書画像の分類を示した図である。

【図8】PICT部とTEXT部の重なりのパターンを説明する図である。

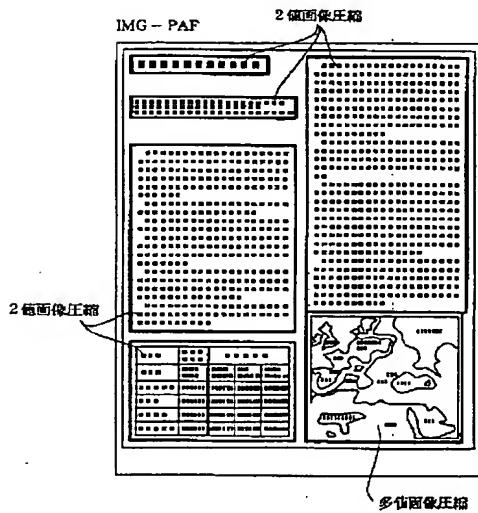
【図9】PICT部とTEXT部の重なりのパターンを説明する図である。

【図10】本実施形態による画像理解処理の手順を示すフローチャートである。

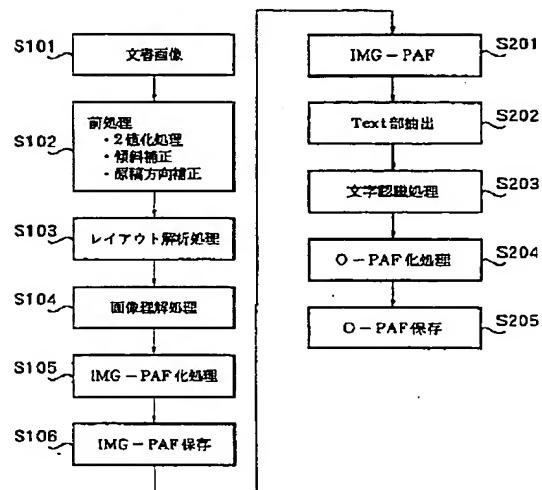
【図11】PICT領域とTEXT領域の結合を説明する図である。

30 【図12】本実施形態による文書画像処理を実現するためのシステム構成を示すブロック図である。

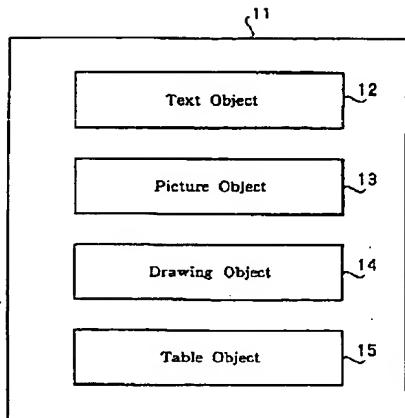
【図2】



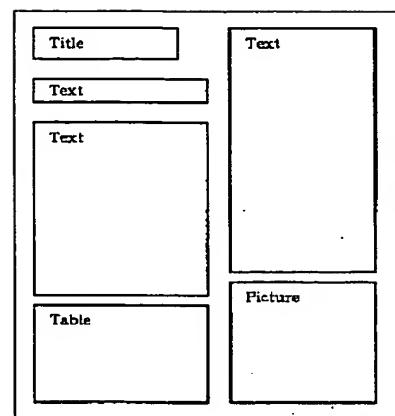
【図3】



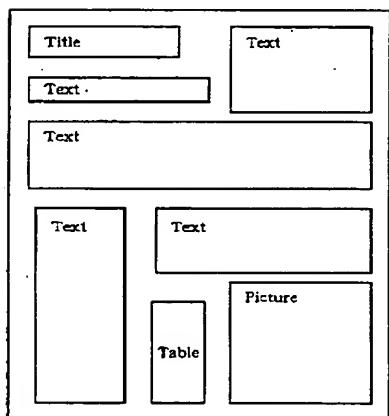
【図1】



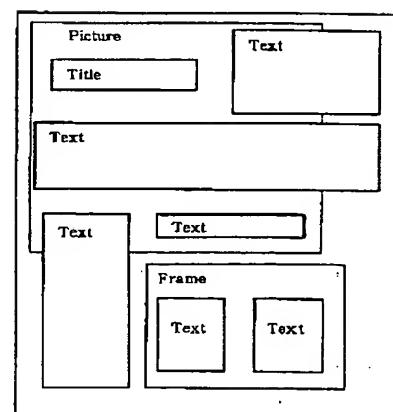
【図4】



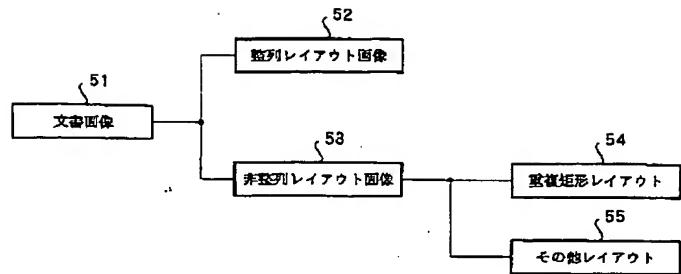
【図5】



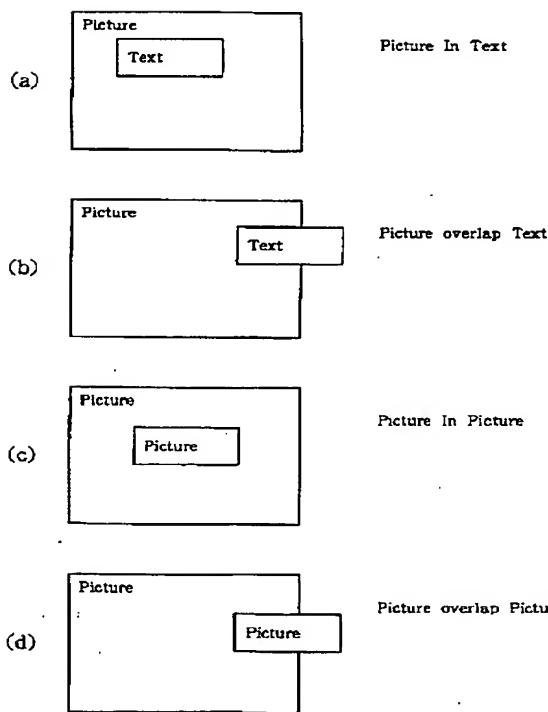
【図6】



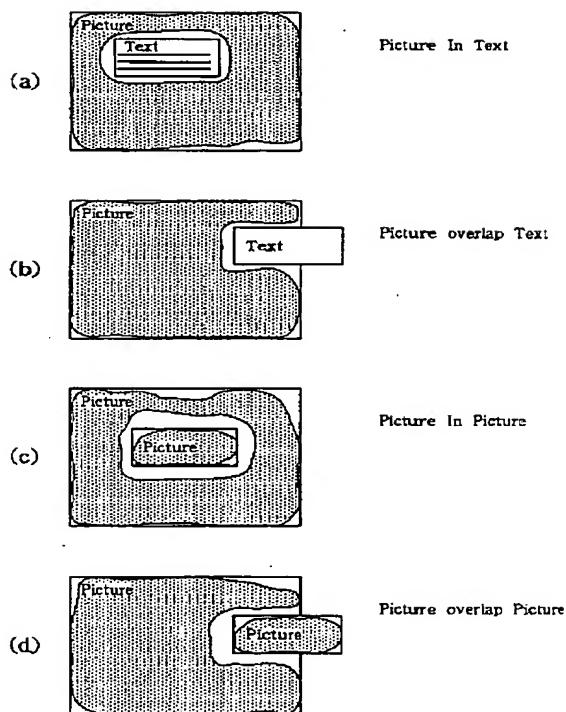
【図7】



【図8】

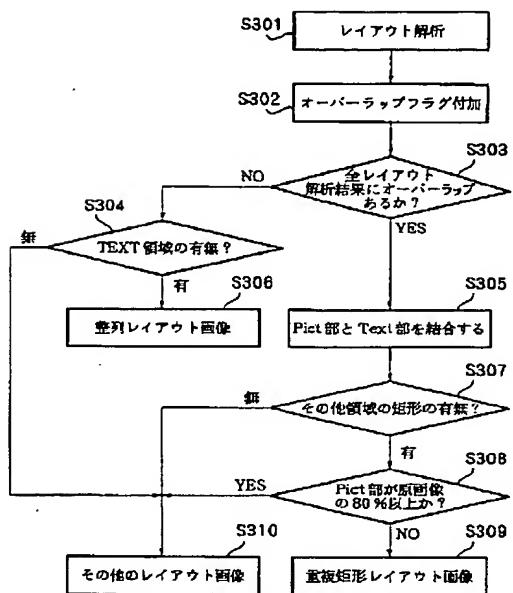


【図9】

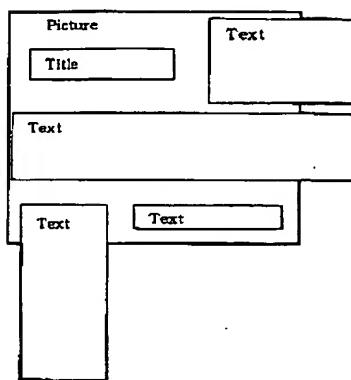


REST AVAILABLE 60PY

【図10】



【図11】



【図12】

